Sentiment Analysis on Textual Data - *A comparison of accuracy using different algorithms*

[1]Ashish Modi, [2]Nambiar Anusree Unnikrishnan
[1]Assistant Professor, [2]Student
[1]Department of Computer and Information Science,
[1]Nagindas Khandwala College, Mumbai, India.
[1]ashishmodi@nkc.ac.in, [2]anusreenambiar2003@gmail.com

***Abstract*— Sentiment Analysis and Opinion Mining is an adverse area of research that deals with the judgement and evaluation criteria in an Emotional communication. Sentiment analysis is generally classified into three different level namely they are document level, sentence level and entity-aspect level. Several research papers are published based on machine learning techniques such as support vector machine, maximum entropy(maxent) and naïve bayes classifiers and random forest classifiers as the most widely used algorithm in sentiment analysis. Our approach is to take dataset and implement the same using three machine learning models and compare their accuracy.**

***Index Terms*— Sentiment analysis, machine learning, Support Vector machine, Naïve bayes classifier, Sentiment polarity.**

I. INTRODUCTION

Sentiment analysis, a branch of natural language processing (NLP), involves extracting and categorizing emotions, opinions, and attitudes from text data. Its relevance in today's data-centric world lies in its ability to interpret unstructured data, providing insights into public perception, consumer sentiment, and market trends. This process empowers businesses to make data-driven decisions, enhance products/services, and understand societal sentiments.

Originating from linguistic analysis, sentiment analysis has evolved from rule-based methods reliant on lexicons to machine learning-based approaches. Early techniques struggled with nuanced sentiments, but advancements in machine learning, including algorithms like Support Vector Machines (SVM) and deep learning models like Transformers, significantly improved accuracy. This evolution enabled the analysis of complex language structures and context, enhancing sentiment classification.

II. LITERATURE SURVEY

(Lee, 2004) In their influential paper titled "Opinion Mining and Sentiment Analysis," Pang and Lee provide a comprehensive overview of sentiment analysis. They explore different approaches, challenges, and applications in sentiment analysis, emphasizing machine learning techniques for classification. This paper laid the groundwork for sentiment analysis research, highlighting the significance of the field and discussing various methodologies that subsequent studies built upon.

(Andrea Esuli, 2006) They introduced "SentiWordNet," a publicly available lexical resource that assigns sentiment scores to words. SentiWordNet facilitated sentiment analysis based on semantic word-level information. This resource provided a valuable foundation for sentiment analysis by offering a comprehensive list of words with associated sentiment scores, enabling more nuanced sentiment analysis based on word semantics.

(Hutto, 2014) VADER is a lexicon and rule-based approach specifically tailored for sentiment analysis of social media text. VADER became widely used in sentiment analysis tasks involving social media due to its ability to handle informal language and context, providing accurate sentiment scores for short texts.

(Toutanova) While not solely focused on sentiment analysis, BERT's contextual understanding capabilities significantly impacted sentiment analysis tasks. BERT revolutionized NLP tasks, including sentiment analysis, by pre-

training on vast amounts of text data and achieving state-of-the-art performance due to its ability to capture contextual information.

III. RESEARCH METHODOLOGY

The data collection is the first step in sentiment analysis. Data collection for sentiment analysis involves gathering text data from various sources to build datasets for analysis. In this project, the collected data is considered as secondary data because for analysis the data is collected from Kaggle (Kaggle, n.d.) which includes cleaning, normalization (lowercasing, removing special characters), tokenization, and possibly stemming or lemmatization. For sentiment analysis, labeling or categorization of text data based on sentiment is crucial for training machine learning models.

In proposed paper, a comparative study is done using three machine learning algorithms viz., Naïve Bayes, Logistics Regression and Support Vector Machine.

Support Vector Machine (SVM): The Support Vector Machine (SVM) algorithm is a versatile and powerful tool in machine learning used for classification and regression tasks. It begins with a labeled dataset, where each data point has an associated class label or numerical value. The algorithm selects the appropriate kernel, which can be linear or non linear, to transform the data into a higher-dimensional space if needed.

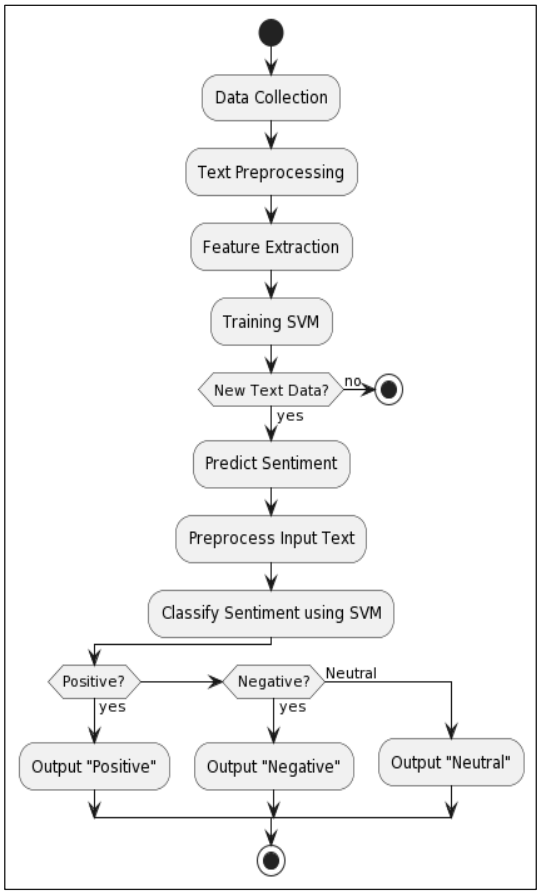

Fig1: Working of Sentiment Analysis using Naïve Bayes Algorithm

Logistic Regression: Logistic Regression is a fundamental machine learning algorithm primarily employed for binary classification tasks. Its operation starts with the preparation of a labeled dataset containing feature variables and a binary target variable (0 or 1). It calculates a linear combination of the features and these coefficients, followed by applying the logistic function, which maps the result to a probability between 0 and 1.

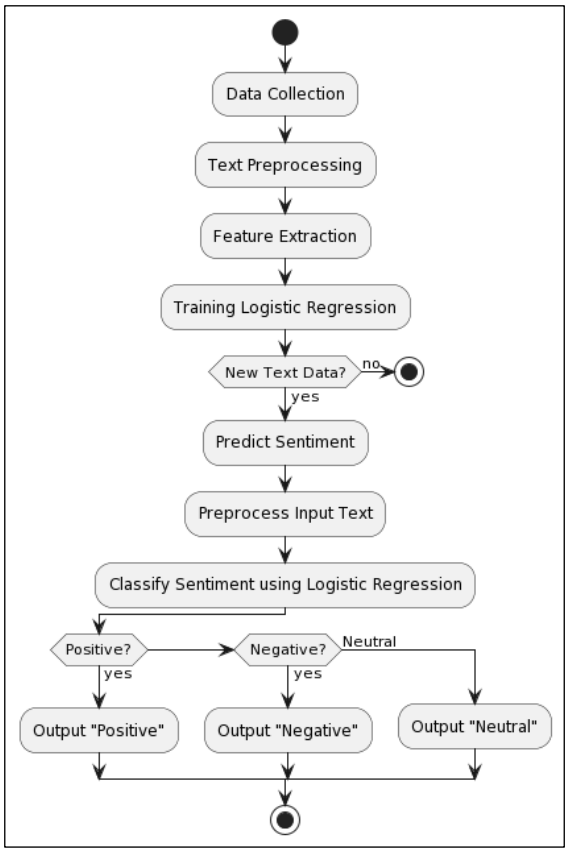

Fig2: Working of Sentiment Analysis using Logistics Regression Algorithm

Naïve Bayes: The Naive Bayes algorithm is a popular and simple machine learning algorithm used for classification tasks, particularly in natural language processing (NLP) and text classification. It's based on Bayes' theorem, which is a probability theory that describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

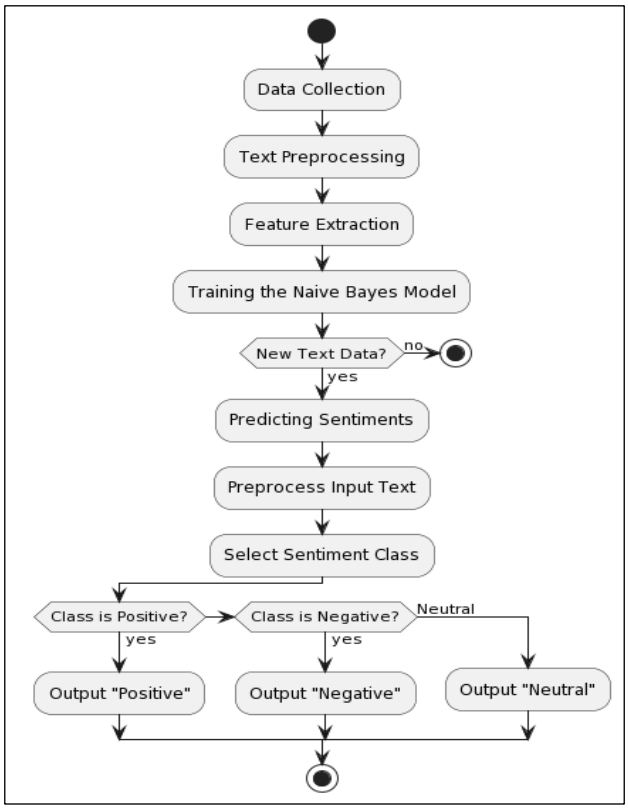

Fig3: Working of Sentiment Analysis using Naïve Bayes Algorithm

IV. FINDINGS

Naive Bayes: In our sentiment analysis study, Naive Bayes algorithm achieving an accuracy rate of 0.76 in accurately predicting sentiments from textual data. SVM: In our sentiment analysis study, Support Vector Machine algorithm achieving an accuracy rate of 0.87 in accurately predicting sentiments from textual data. Logistic Regression: In our sentiment analysis study, Logistic Regression algorithm achieving an accuracy rate of 0.85 in accurately predicting sentiments from textual data.

In our sentiment analysis study, when evaluating the performance of various machine learning models, Support Vector Machine (SVM) demonstrated the highest accuracy at 87%, outperforming Logistic Regression, which achieved an accuracy of 85%. Naive Bayes, while slightly lower in accuracy at 76%, also showcased notable performance in sentiment classification tasks.

V. CONCLUSION

In this study, we explored the efficacy of various machine learning models in sentiment analysis. Our investigation revealed notable accuracies, with Support Vector Machine (SVM) exhibiting the highest accuracy at 87%, closely followed by Logistic Regression at 85%. Naive Bayes demonstrated respectable performance with an accuracy of 76%. These findings emphasize the potential and versatility of machine learning algorithms in deciphering sentiments from textual data.

In conclusion, our research underscores the effectiveness of machine learning models, particularly SVM and Logistic Regression, in accurately discerning sentiments from textual data. These findings pave the way for informed decision-making in diverse domains and provide a foundation for further advancements in sentiment analysis methodologies.

REFERENCES

(N.D.). RETRIEVED FROM KAGGLE: HTTPS://WWW.KAGGLE.COM/GPREDA/PFIZER-VACCINE-TWEETS

Andrea Esuli, F. S. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06).* Genoa, Italy: European Language Resources Association (ELRA).

Hutto, C. &. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference on Weblogs and Social Media.*

Lee, B. P. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, (pp. 271–278). Barcelona, Spain.

Toutanova, J. D.-W. (n.d.). BERT: Pre-training of Deep Bidirectional Transformers for.